Working paper
2020-04
Statistics and Econometrics
ISSN 2387-0303

Iterative variable selection for high-dimensional data: Prediction of pathological response in triple-negative breast cancer

Juan C. Laria, M. Carmen Aguilera-Morillo, Enrique Álvarez, Rosa E. Lillo, Sara López-Taruella, María del Monte-Millán, Antonio C. Picornell, Miguel Martín, Juan Romo

uc3m | Universidad Carlos III de Madrid

ITERATIVE VARIABLE SELECTION FOR HIGH-DIMENSIONAL DATA: PREDICTION OF PATHOLOGICAL RESPONSE IN TRIPLE-NEGATIVE BREAST CANCER

WORKING PAPER

Juan C. Laria
Department of Statistics,
University Carlos III of Madrid
UC3M-BS Institute of Financial Big Data
juancarlos.laria@uc3m.es

M. Carmen Aguilera-Morillo
Department of Applied Statistics and
Operational Research and Quality,
Universitat Politècnica de València
UC3M-BS Santander Big Data Institute

Enrique Álvarez
Department of Medical Oncology,
Hospital General Universitario
Gregorio Marañón

Rosa E. Lillo
Department of Statistics,
University Carlos III of Madrid
UC3M-BS Institute of Financial Big Data

Sara López-Taruella
Department of Medical Oncology,
Hospital General Universitario
Gregorio Marañón

María del Monte-Millán
Department of Medical Oncology,
Hospital General Universitario
Gregorio Marañón

Antonio C. Picornell
Department of Medical Oncology,
Hospital General Universitario
Gregorio Marañón

Miguel Martín
Department of Medical Oncology,
Hospital General Universitario
Gregorio Marañón

Juan Romo
Department of Statistics,
University Carlos III of Madrid
UC3M-BS Institute of Financial Big Data

June 4, 2020

ABSTRACT

In the last decade, regularized regression methods have offered alternatives for performing multi-marker analysis and feature selection in a whole genome context. The process of defining a list of genes that will characterize an expression profile, remains unclear. This procedure oscillates between selecting the genes or transcripts of interest based on previous clinical evidence, or performing a whole transcriptome analysis that rests on advanced statistics. This paper introduces a methodology to deal with the variable selection and model estimation problems in the high-dimensional set-up, which can be particularly useful in the whole genome context. Results are validated using simulated data, and a real dataset from a triple-negative breast cancer study.

***K*eywords** variable selection, high-dimension, regularization, classification

1 Introduction

Breast cancer (BC) is the most frequent cancer among women, representing around 25% of all newly diagnosed cancer in women (Ferlay et al., 2014). One in eight women in developed countries will be diagnosed with BC over the course of a lifetime.

The prognosis of this disease has progressively improved over the past three decades, due to the implementation of population-based screening campaigns and, above all, the introduction of new effective targeted medical therapies, i.e., aromatase inhibitors (effective in hormone receptor-positive tumors) and trastuzumab (effective in HER2-positive tumors). Breast cancer is, however, a heterogeneous disease. The worst outcomes are associated with the so-called triple-negative breast cancer subtype (TNBC), diagnosed in 15-20% of BC patients. TNBC is defined by a lack of immunohistochemistry expression of the estrogen and progesterone receptors and a lack of expression/amplification of HER2 (Dent et al., 2007). The absence of expression of these receptors makes chemotherapy the only available therapy for TNBC.

TNBC is usually diagnosed in an operable (early) stage. Surgery, chemotherapy and radiation therapy are the critical components of the treatment of early TNBC. Many early TNBC patients are treated with upfront chemotherapy (neoadjuvant chemotherapy, NACT) and then operated on and, perhaps, irradiated. The rationale for this sequence is the ability to predict the long-term outcome of patients looking at the pathological response achieved with initial NACT (Cortazar et al., 2014). With the currently available neoadjuvant chemotherapy regimens, nearly 50% of TNBC achieve a good pathological response to this therapy, while the remaining patients have an insufficient response. TNBC patients achieving a complete or almost complete disappearance of the tumor in breast and axilla after NACT have an excellent outcome (less than 10% of relapses at five years), in contrast with those with significant residual disease (more than 50% of relapses at five years) (Symmans et al., 2017; Sharma et al., 2018).

The identification of these two different populations is therefore of the utmost relevance, in order to test new experimental therapies in the population unlikely to achieve a good pathological response. Several tumor multigene predictors of pathological response of operable BC to NACT have been proposed in the past few years, taking advantage of the recent decreased economic cost of obtaining an individual's full transcriptome (Tabchy et al., 2010; Hatzis et al., 2011; Chang et al., 2003). Most of them have been tested in unselected populations of BC patients and have shown insufficient positive predictive value and sensitivity.

The process of defining a list of genes that will define a characteristic expression profile is still ambiguous. This process oscillates between selecting the genes or transcripts of interest based on the clinical evidence in previous studies or using an agnostic point of view that rests on advanced statistics selection processes in multivariate analysis. RNA-Seq has become one of the most appealing tools of modern whole transcriptome analyses because it combines relatively low cost and a comprehensive approach to transcript quantification. Some approaches to complex disease biomarker discovery already pointed to the need to use a whole genome perspective using joint information in order to predict complex traits instead of a priori selecting individual features (De Los Campos et al., 2010; Lupski et al., 2011). This strategy would lead to high predictive accuracy, and there would be no need to know the precise biological associations in the genome background because of the high correlation among the biomarkers (Offit, 2011). This approach is challenging from the statistical point of view because of the large number of biomarkers to be tested along the genome related to the rather small sample sizes in clinical studies. On the other hand, daily clinical practice scenario requires cheaper and faster quantification platforms than whole-genome RNA-Seq analysis. Thus, it is needed to reduce the number of biomarkers to stick with in order to define a practical gene expression signature for the clinical community.

The regularized regression methods provide alternatives for performing multi-marker analysis and feature selection in a whole genome context (Szymczak et al., 2009). Specifically, we focus on the sparse-group lasso (SGL) regularization method (Simon et al., 2013), which generalizes lasso (Tibshirani, 1996), group lasso (Yuan and Lin, 2006) and elastic-net (Zou and Hastie, 2003), merging lasso and group lasso penalties. The solution provided by SGL, usually involves a small number of predictor variables, given that many coefficients in the solution are exactly zero. It has an advantage over lasso when the predictor variables are grouped, as many groups are entirely zeroed out, but unlike group lasso, the solution is also sparse within those groups that are not completely eliminated from the model. However, as will be explained in next sections, the SGL is not appropriate for the problem we are dealing with, without introducing a broader methodology to control the regularization hyper-parameters, the groups, and the high-dimensionality issue.

From a methodological point of view, this paper provides an original contribution to perform variable selection and model fitting in high-dimensional problems. Furthermore, the results presented in this paper are the first attempt in a Translational Oncology scenario of building a predictive model for the response to treatment, based entirely on the whole genome RNA-Seq data and conventional clinical variables.

This paper is organized as follows. Section 2 ties together the various theoretical concepts that support our approach. Section 2.1 introduces the mathematical formulation of the SGL, as an optimization problem. Section 2.2 discusses the iterative-sparse group lasso, a coordinate descent algorithm to automatically select the regularization parameters of the SGL. Section 2.3 describes a clustering strategy for the variables, based on principal component analysis, which makes it possible to work with an arbitrarily large number of variables, without specifying the groups apriori. Section 2.5 highlights our main methodological contributions: the *importance* and the *power indexes*, to weight variables and models, respectively. In Section 3, a simulation study is presented, with several synthetic matrix designs, and varying the number of variables from 40 to 4000. Section 4 highlights the contributions of our methodology on a TNBC cohort which had undergone neoadjuvant docetaxel/carboplatin chemotherapy. Some conclusions and lines for future work, are drawn in the final section.

2 Methodology and algorithms

Consider the usual logistic regression framework, with N observations in the form $\{y^{(i)}, x_1^{(i)}, x_2^{(i)}, \ldots, x_p^{(i)}\}_{i=1}^N$, where p is the number of features or predictor variables, and $y^{(i)}$ is the binary response. We assume that the response comes from a random variable with conditional distribution,

$$Y|(X_1 \ldots X_p) \sim Ber(p(X_1 \ldots X_p, \boldsymbol{\beta})),$$

where

$$p(X_1 \ldots X_p, \boldsymbol{\beta}) = (1 + \exp(-\eta))^{-1},$$

and η is the linear predictor,

$$\eta = \beta_0 + \sum_{j=1}^{p} \beta_j X_j, \quad \boldsymbol{\beta} = [\beta_0\, \beta_1\, \ldots\, \beta_p] \in \mathbb{R}^{p+1}.$$

The objective is to predict the response Y for future observations of $X_1 \ldots X_p$, using an estimation of the unknown parameter $\boldsymbol{\beta}$, given by,

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta} \in \mathbb{R}^{p+1}}{\operatorname{argmin}} \hat{R}(\boldsymbol{\beta}), \tag{1}$$

were

$$\hat{R}(\boldsymbol{\beta}) = \frac{1}{N} \sum_{i=1}^{N} \left[\log \left(1 + \exp\{\beta_0 + \sum_{j=1}^{p} \beta_j x_j^{(i)}\} \right) - y_i (\beta_0 + \sum_{j=1}^{p} \beta_j x_j^{(i)}) \right]. \tag{2}$$

The problem with this approach is that for $N < p$, the minimization (1) has infinite optimal solutions. When the features $X_1 \ldots X_p$ represent genetic expressions, this problem of predicting Y becomes more extreme, since we often have N several orders of magnitude smaller than p.

As a solution, variable selection techniques are proposed, in order to tackle the analytical intractability of this problem.

2.1 The sparse-group lasso

It has been shown that SGL can play an important role in addressing the issue of variable selection in genetic models, where genes are grouped following different pathways. The mathematical formulation of this problem is,

$$\hat{\boldsymbol{\beta}}(\boldsymbol{\lambda}) = \underset{\boldsymbol{\beta} \in \mathbb{R}^{p+1}}{\operatorname{argmin}} \left\{ \hat{R}(\boldsymbol{\beta}) + \lambda_2 \sum_{j=1}^{J} \gamma_j \|\boldsymbol{\beta}^{(j)}\|_2 + \lambda_1 \|\boldsymbol{\beta}\|_1 \right\}. \tag{3}$$

Here J is the number of groups, and $\boldsymbol{\beta}^{(j)} \in \mathbb{R}^{p_j}$ are vectors with the components of $\boldsymbol{\beta}$ corresponding to j-th group (of size p_j), and $\gamma_j = \sqrt{p_j}$, $j = 1, 2, \ldots, J$. The regularization parameter is $\boldsymbol{\lambda} = [\lambda_1\, \lambda_2] \in \mathbb{R}_+^2$.

The problem with (3) is that vector $\hat{\boldsymbol{\beta}}(\boldsymbol{\lambda})$ of estimated coefficients depends on the selection of a vector of regulation parameters $\boldsymbol{\lambda}$, which must be chosen before estimating $\hat{\boldsymbol{\beta}}(\boldsymbol{\lambda})$. The selection of $\boldsymbol{\lambda}$ is partly an open problem, because although there are several practical strategies for choosing these parameters, there is no established theoretical criterion to follow. In most cases, the regularization parameters are set a priori, based on some additional information about the data, or the

characteristics of the desired solution, e.g., greater λ_1 implies more components of $\hat{\boldsymbol{\beta}}$ identically zero. The most commonly used methodology to select $\boldsymbol{\lambda}$ consists of moving the regulation parameters in a fixed grid, usually not very thin. However, this approach has many disadvantages.(Laria et al., 2019) In contrast, we propose the iterative-sparse group lasso, a coordinate descent algorithm, recently introduced by Laria et al..

2.2 Selection of the optimal regularization parameter

Traditionally, the data set $\mathcal{Z} = \{y^{(i)}, x_1^{(i)}, x_2^{(i)}, \ldots, x_p^{(i)}\}_{i=1}^N$ is partitioned into three disjoint data sets, $\mathcal{Z}_T, \mathcal{Z}_V$ and $\mathcal{Z}_{test}$. The data in $\mathcal{Z}_T$ is used for training the model, i.e., solving (3). $\mathcal{Z}_V$ is used for validation, i.e., finding the optimal parameter $\boldsymbol{\lambda}$. The remaining observations in $\mathcal{Z}_{test}$ are used for testing the prediction ability of the model on future observations. Specifically, the selection of the optimal parameter $\boldsymbol{\lambda}$ is based on the minimization of the validation error, defined as

$$\hat{R}_V(\boldsymbol{\lambda}) = \frac{1}{\#\mathcal{Z}_V} \sum_{(y^{(i)}, \mathbf{x}^{(i)}) \in \mathcal{Z}_V} \left[\log(1 + \exp\{\eta(\hat{\boldsymbol{\beta}}_T)\}) - y^{(i)} \eta(\hat{\boldsymbol{\beta}}_T) \right], \tag{4}$$

where

$$\hat{\boldsymbol{\beta}}_T(\boldsymbol{\lambda}) = \arg\min_{\boldsymbol{\beta} \in B} \left\{ \hat{R}_T(\boldsymbol{\beta}) + \lambda_2 \sum_{j=1}^{J} \gamma_j \|\boldsymbol{\beta}^{(j)}\|_2 + \lambda_1 \|\boldsymbol{\beta}\|_1 \right\}, \tag{5}$$

and

$$\hat{R}_T(\boldsymbol{\beta}) = \frac{1}{\#\mathcal{Z}_T} \sum_{(y^{(i)}, \mathbf{x}^{(i)}) \in \mathcal{Z}_T} \left[\log(1 + \exp\{\eta(\hat{\boldsymbol{\beta}}_T)\}) - y^{(i)} \eta(\hat{\boldsymbol{\beta}}_T) \right], \tag{6}$$

with $\#$ denoting the cardinal of a set. Therefore, the problem of finding the optimal parameter $\boldsymbol{\lambda}$ can be formulated as,

$$\begin{aligned} &\min_{\boldsymbol{\lambda} \in \mathbb{R}_+^2} \hat{R}_V(\boldsymbol{\lambda}) \\ \text{s.t.} \quad &\hat{\boldsymbol{\beta}}_T(\boldsymbol{\lambda}) = \arg\min_{\boldsymbol{\beta} \in \mathbb{R}^{p+1}} \left\{ \hat{R}_T(\boldsymbol{\beta}) + \lambda_2 \textstyle\sum_{j=1}^{J} \gamma_j \|\boldsymbol{\beta}^{(j)}\|_2 + \lambda_1 \|\boldsymbol{\beta}\|_1 \right\}. \end{aligned} \tag{7}$$

Algorithm 1 describes the two-parameter ITERATIVE SPARSE-GROUP LASSO (iSGL$_0$), a gradient-free coordinate descent method to tune the parameter $\boldsymbol{\lambda}$ from the sparse-group lasso (3), which performs well under different scenarios while drastically reducing the number of operations required to find optimal penalty weight parameters that minimize the validation error in (4). The iSGL$_0$ iteratively performs a univariate minimization over one of the coordinates of $\boldsymbol{\lambda}$, while the other coordinate is fixed.

Algorithm 1: TWO-PARAMETER ITERATIVE SPARSE-GROUP LASSO (iSGL$_0$)

```
/* Data for training/validation                                    */
Function isgl(Z_T, Z_V):
    Initialize λ  i ← 1
    while λ not stationary do
        λ_i ← argmin_{λ∈R_+} R̂_V(λ | λ_i = λ);   // minimize over coordinate i of λ
        i ← i mod 2 + 1;                          // Next coordinate
    end
    return β̂_T(λ)
```

Laria et al. (2019) provide detailed information about Algorithm 1 in their paper. As mentioned before, a very useful property of the sparse-group lasso as a variable selection method, is the ability to remove entire groups from the model (sending to zero the components of the $\hat{\boldsymbol{\beta}}$ vector relative to those groups), as is the case with group lasso. However, this means that a grouping among the variables under consideration must be specified. This does not entail a challenge if there are natural groupings among the variables, for example, if the variables are dummies related to different levels of the same original categorical variable. However, in our study most of the variables are transcriptomes, for which there are no established groupings in the literature. To overcome this problem, we suggest an empirical variable grouping approach, based on the principal component analysis of the data matrix.

2.3 Grouping variables using principal component analysis

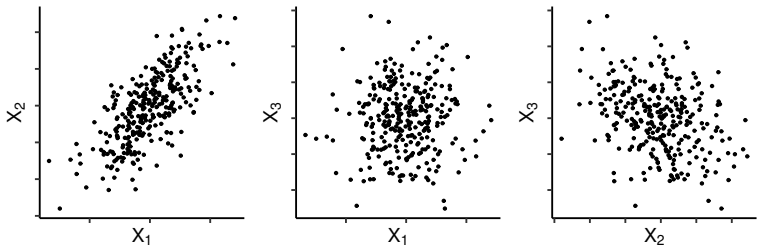
Principal component analysis (PCA) is a dimension reduction technique, very effective in reducing a large number of variables related to each other to a few latent variables, trying to lose the minimum amount of information. The new latent variables obtained (the principal components), which are a linear transformation of the original variables, are uncorrelated and ordered in such a way that the first components capture most of the variation present in all the original variables.

Given the data matrix $\mathbf{X} \in \mathbb{R}^{N\times p}$, PCA computes the rotation matrix $\mathbf{W} \in \mathbb{R}^{p\times G}$, where $G \leq \min(N, p)$ is the number of principal component to retain. The transformed data matrix (the principal component matrix) is $\mathbf{T} = \mathbf{XW}$. This rotation matrix $\mathbf{W}$ suggest a natural grouping on the columns of $\mathbf{X}$, given by

$$group(X_j) = \arg\max_i |W_{ji}|, \quad j = 1, 2, \ldots, p. \tag{8}$$

This strategy will provide at most G groups on the columns of $\mathbf{X}$.

Figure 1: Simulated sample from three random variables, that illustrate the grouping based on PCA.



The following example illustrates our approach on a simulated data set. Suppose that we want to cluster variables X_1, X_2 and X_3 using two groups. There are 300 observations (Fig. 1) and they are simulated such that $\text{corr}(X_1, X_2) = 0.75$, $\text{corr}(X_1, X_3) = 0.1$ and $\text{corr}(X_2, X_3) = -0.25$. The principal component's rotation matrix $\mathbf{W}$ is given by,

	PC1	PC2
X_1	**-0.67**	0.40
X_2	**-0.70**	-0.08
X_3	0.23	**0.91**

In this example, X_1 and X_2 would be grouped together, whereas X_3 would be in the other group. Apparently, this method is placing highly correlated variables in the same group.

2.4 Mining influent variables under a cross-validation approach

In this section, we focus on the problem of variable selection in models where the ratio p/N is in the order of 10^2. In these scenarios, even state-of-the-art methods such as SGL find it hard to select an appropriate set of variables related to the response term. We propose a cross-validation approach to fit and evaluate many different models using only a sample size of N observations initially given.

The solution in terms of $\hat{\boldsymbol{\beta}}(\boldsymbol{\lambda})$ provided by Algorithm 1 strongly depends on the partition $\mathcal{Z}_T, \mathcal{Z}_V$. As a consequence, if we run Algorithm 1 for different partitions $\mathcal{Z}_T, \mathcal{Z}_V$ of the same data $\mathcal{Z}$, it will probably result in different coefficient estimates $\hat{\boldsymbol{\beta}}(\boldsymbol{\lambda})$. Therefore, the indicator function of variable X_j included in the model, $\mathbb{I}(\hat{\beta}_j(\boldsymbol{\lambda}) \neq 0)$, will take different values depending on the partition $\mathcal{Z}_T, \mathcal{Z}_V$. In order to avoid this dependency on the sample data partition, we propose Algorithm 2, which computes many different solutions $\hat{\boldsymbol{\beta}}(\boldsymbol{\lambda})$ of Algorithm 1, for different partitions of the original data sample $\mathcal{Z}$. The goal of this algorithm is to be able to fit and evaluate many models using the same data. Since the sample size is small compared to the number of covariates, the variable selection will greatly depend on the train/validate partition. We denote by R the total number of models that will be fitted using different partitions from the original sample. Algorithm 2 stores the information of the fitting $\hat{\boldsymbol{\beta}}$ of each model and the correct classification rate in the validation sample (ccr_V) in each case.

Algorithm 2:

```
/* sample data Z, # of runs R                                    */
```
Function `isgl(`$\mathcal{Z}$, R`)`**:**
  for r *in* $1, 2 \dots R$ **do**
    $\mathcal{Z}_T, \mathcal{Z}_V \leftarrow$ random partition of $\mathcal{Z}$
    $\boldsymbol{\beta}^{(r)} \leftarrow$ ISGL(Z_T, Z_V)
    $ccr_V^{(r)} \leftarrow$ Correct classification rate of $\boldsymbol{\beta}^{(r)}$ in $\mathcal{Z}_V$
  end
  return $\boldsymbol{\beta}$, ccr_V

2.5 Selection of the best model

Our objective is to select one of those R models computed in Algorithm 2 to be our final model. We believe that a selection only based on the maximization of ccr_V could lead to overfit in the training sample data $\mathcal{Z}$. To overcome this problem, we define two indexes: the *importance index* of a variable, and the *power* of a model. These indexes are fundamental to choosing a final model that is not overfitting the data.

We consider the *importance index* I_j of variable X_j defined as,

$$I_j = \sum_{r=1}^{R} |\beta_j^{(r)}| \cdot (ccr_V^{(r)} - \delta) / \max_j \left\{ \sum_{r=1}^{R} |\beta_j^{(r)}| \cdot (ccr_V^{(r)} - \delta) \right\}, \tag{9}$$

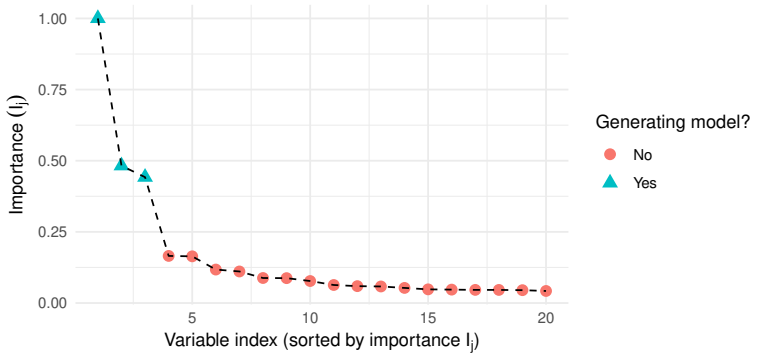
where $\boldsymbol{\beta}^{(r)}$ and $ccr_V^{(r)}$ are those returned by Algorithm 2 on the data $\mathcal{Z}$. With the objective of penalizing those models that had a bad performance on the validation set, the term δ has been

introduced, which is the maximum between $\bar{\boldsymbol{y}}$ and $1 - \bar{\boldsymbol{y}}$, i.e., the null model correct classification rate.

The *importance index* weights differently each variable $X_1 \dots X_p$ depending on the correct classification rate of those models in which each variable was present. The larger I_j, the greater the chances of X_j being present in the underlying model that generated the data $\mathcal{Z}$.

Figure 2 illustrates the *importance index*, computed on a simulated data set, with $N = 100$ observations and $p = 400$ variables. Notice that the highest three variables in importance are actually in the generating model, and there is a clear gap in Fig. 2 between them and the rest of the variables.

Figure 2: Sorted *importance index* obtained from Algorithm 2, with $R = 150$, and a simulated data sample with $N = 100$ observations and $p = 400$ variables.



Based on the maximization of the *importance index*, an appropriate subset is selected from the original p variables. Although the true number of variables involved in the model is unknown, we can focus our attention on a predefined number of important variables K, which depends only on the sample data $\mathcal{Z}$. We empirically found $K = \lceil \sqrt{N/2} \rceil$ to achieve good results. Using the important index of the best K variables, we define the *power* of a model as,

$$P_r = \frac{1}{\sum_{k=1}^{K} I_{(k)}} \sum_{j: I_j \leq I_{(K)}} I_j |\beta_j^{(r)}| / \|\boldsymbol{\beta}^{(r)}\|_1, \quad r = 1, 2 \dots R, \tag{10}$$

where $I_{(k)}$ denotes the $k-$th greatest *importance index*, e.g., $I_{(1)} = \max_j I_j$. The *power index* P weights each model, depending on the *importance* of its included variables.

The selection of the final model is based on the criterion,

$$\hat{\boldsymbol{\beta}} = \boldsymbol{\beta}^{(r^*)}, \quad \text{where} \quad r^* = \max_r \left\{ P_r + ccr_V^{(r)} \right\}. \tag{11}$$

Equation (11), Algorithm 2, and the framework that supports them, is the main contribution of this paper from a methodological point of view. Equation (11) is based on the correct classification rates of R different fitted models, two indexes defined in this paper, and the iterative sparse-group lasso, which is a novel algorithm.

3 A simulation study

In this section, we illustrate the performance of Algorithm 2 using synthetic data. To generate observations, we have followed simulation designs from Simon et al. (2013) (uncorrelated features),

Tibshirani (1996), Zou and Hastie (2005) and Azevedo Costa et al. (2017) (correlated features). Since our objective was to evaluate Algorithm 2 in binary classification problems, we used a logistic regression model for the response term using the simulated design matrices in each case. We simulated data from the true model,

$$\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta},$$

with logistic response $\boldsymbol{y}$ given by

$$y_i \sim Ber(p_i), \quad p_i = (1 + \exp(-\eta_i))^{-1}, \quad i = 1, 2 \dots N. \tag{12}$$

Five scenarios for $\boldsymbol{\beta}$ and $\mathbf{X}$ were simulated. In each example, our simulated data consisted of a training set of $N = 100$ observations and p variables, and an independent test set of 5000 observations and p variables. Models were fitted using training data only. Here are the details of the five scenarios.

SFHT_1) This example is adapted from the sparse-group lasso paper (Simon et al., 2013). We set

$$\boldsymbol{\beta} = (1, 2, 3, 4, 5, \underbrace{0, \dots, 0}_{p-5})$$

and X_i are i.i.d $N(0, 1)$, for $1 \leq i \leq p$.

SFHT_2) In this example, $\boldsymbol{\beta}$ is generated as in SFHT_1, but the rows of the model matrix $\mathbf{X}$ are i.i.d. generated from a multivariate gaussian distribution with $\text{cov}(X_i, X_j) = 0.5^{|i-j|}, 1 \leq j \leq i \leq p$.

Tibs_1) This example is adapted from the original lasso paper (Tibshirani, 1996), also found in other simulation studies (Zou and Hastie, 2005; Azevedo Costa et al., 2017). We set

$$\boldsymbol{\beta} = (3, 1.5, 0, 0, 2, \underbrace{0, \dots, 0}_{p-5}),$$

and the rows of $\mathbf{X}$ are i.i.d. generated from a multivariate gaussian distribution with $\text{cov}(X_i, X_j) = 0.5^{|i-j|}, 1 \leq j \leq i \leq p$.

Tibs_4) This example is also adapted from the original lasso paper (Tibshirani, 1996), and found in other simulation studies as well (Zou and Hastie, 2005; Azevedo Costa et al., 2017). We set

$$\boldsymbol{\beta} = (\underbrace{0, \dots, 0}_{10}, \underbrace{2, \dots, 2}_{10}, \underbrace{0, \dots, 0}_{10}, \underbrace{2, \dots, 2}_{10}, \underbrace{0, \dots, 0}_{p-40})$$

and the rows of $\mathbf{X}$ are i.i.d. generated from a multivariate gaussian distribution with $\text{cov}(X_i, X_j) = 0.5$, and $\text{var}(X_i) = 1,\ 1 \leq j < i \leq p$.

ZH_d) This example is adapted from the elastic net paper (Zou and Hastie, 2005). We chose

$$\boldsymbol{\beta} = (\underbrace{3, \dots, 3}_{15}, \underbrace{0, \dots, 0}_{p-15})$$

and the rows of $\mathbf{X}$ were generated as follows,

$$X_i = Z_1 + \epsilon_i^x, \quad Z_1 \sim N(0, 1), \quad i = 1, \dots, 5,$$

$$X_i = Z_2 + \epsilon_i^x, \quad Z_2 \sim N(0,1), \quad i = 6, \ldots, 10,$$
$$X_i = Z_3 + \epsilon_i^x, \quad Z_3 \sim N(0,1), \quad i = 11, \ldots, 15,$$
$$X_i \sim N(0,1),\ X_i \text{ i.i.d. for } i = 16, \ldots, p,$$

where ϵ_i^x are i.i.d. $N(0, 0.01)$, for $1 \leq i \leq 15$.

We aimed to investigate the robustness of our methodology in each example, regarding several measures, as the number of noisy variables (not in the generating model) increased. The criteria we used to evaluate the models in each case were the correct classifications rate in the test sample (ccr), the correct classifications rate in the training sample $\mathbb{E}_t(ccr)$, and the specificity ($spec.$) and sensitivity ($sens.$) concerning variable selection. Let $\hat{\boldsymbol{\beta}}$ be the final estimated coefficient vector and $\boldsymbol{\beta}$ the true generating coefficient vector, then the sensitivity was measured as

$$sens. = \sum_{j=1}^{p} \mathbb{I}(\hat{\beta}_j \neq 0) \cdot \mathbb{I}(\beta_j \neq 0) / \sum_{j=1}^{p} \mathbb{I}(\hat{\beta}_j \neq 0)$$

Analogously, the specificity was defined as

$$spec. = \sum_{j=1}^{p} \mathbb{I}(\hat{\beta}_j = 0) \cdot \mathbb{I}(\beta_j = 0) / \sum_{j=1}^{p} \mathbb{I}(\hat{\beta}_j = 0).$$

Table 1 describes the performance of the final model selected under our methodology in the scenarios described above. We have conducted 30 experiments in each case, as we varied the number of variables in the model (p). Standard deviations are given in parenthesis. Table 1 reveals that for all the configurations (except, perhaps SFHT_1) the methodology is very robust with respect to an increase in the number of variables p. In fact, for most of them, the ccr does not vary much from $p = 400$ to $p = 4000$. Intuitively, the grouping strategy introduced in Section 2.3 places highly correlated variables in the same groups, producing better results when there is correlation between the variables in the model. That is why the simulation scheme SFHT_1 produces the poorest results. In SFHT_1, all the simulated variables are independent and therefore, there is not any clear way to group the variables.

4 Application to Biomedical Data

In this section, we evaluate the methodology described in Algorithm 2 with the model selection criterion given by (11) on a real case study. A sample of TNBC patients from a previously published clinical trial (Sharma et al., 2016) was used to analyze relations between cancer cells transcriptome and the response of patients to the given medical treatment (docetaxel plus carboplatin). The dataset was composed of 93 observations (patients) and 16616 variables (genetic transcripts and clinical variables).

Figure 3 shows the highest 30 *importance indexes* out of a total of 16616 variables. The criterion to measure the importance of the variables is given in (9). Algorithm 2 was run with $R = 200$, and the cutoff value was set to $K = \lceil \sqrt{N/2} \rceil = 7$, as described in Section 2.5. With this importance index, the power of each model was computed using (10) and the best model was chosen according to (11), as highlighted in Fig. 4.

The selected model included 843 out of 16616 variables. The grouping strategy commented in Section 2.3 found a total of 82 groups, from which 18 were included in the final model.

Table 1: Average correct classification rate (ccr) of the final model in the test data set (5000 observations), in 30 experiments for each configuration. $E_t(ccr)$ denotes the estimated correct classification rate from the training sample. The mean sensitivity ($sens.$) and the specificity ($spec.$) with respect to variable selection are also given. Standard deviations are given in parenthesis. Algorithm 2 was run with $R = 200$, and $N = 100$ observations in the training sample.

		Number of variables in the model (p)				
Case		40	100	400	1000	4000
SFHT_1	ccr	**0.84** (0.03)	**0.80** (0.04)	**0.76** (0.04)	**0.73** (0.05)	**0.66** (0.06)
	$E_t(ccr)$	0.90 (0.04)	0.87 (0.05)	0.86 (0.05)	0.81 (0.04)	0.80 (0.05)
	$sens.$	0.83 (0.16)	0.71 (0.24)	0.65 (0.18)	0.59 (0.18)	0.47 (0.17)
	$spec.$	0.66 (0.14)	0.83 (0.09)	0.93 (0.04)	0.96 (0.03)	0.98 (0.02)
SFHT_2	ccr	**0.87** (0.02)	**0.86** (0.02)	**0.84** (0.04)	**0.83** (0.04)	**0.82** (0.04)
	$E_t(ccr)$	0.93 (0.04)	0.94 (0.04)	0.92 (0.04)	0.91 (0.04)	0.90 (0.05)
	$sens.$	0.83 (0.18)	0.78 (0.16)	0.74 (0.16)	0.71 (0.16)	0.61 (0.19)
	$spec.$	0.68 (0.17)	0.80 (0.10)	0.92 (0.04)	0.96 (0.03)	0.99 (0.01)
Tibs_1	ccr	**0.82** (0.02)	**0.81** (0.04)	**0.79** (0.04)	**0.77** (0.04)	**0.76** (0.04)
	$E_t(ccr)$	0.90 (0.03)	0.90 (0.04)	0.88 (0.05)	0.87 (0.05)	0.85 (0.05)
	$sens.$	0.99 (0.06)	0.98 (0.08)	0.92 (0.14)	0.90 (0.18)	0.81 (0.19)
	$spec.$	0.68 (0.14)	0.82 (0.08)	0.92 (0.04)	0.96 (0.02)	0.99 (0.01)
Tibs_4	ccr	**0.91** (0.03)	**0.90** (0.02)	**0.89** (0.01)	**0.90** (0.02)	**0.91** (0.01)
	$E_t(ccr)$	0.97 (0.02)	0.96 (0.03)	0.95 (0.03)	0.87 (0.05)	0.97 (0.02)
	$sens.$	0.71 (0.17)	0.43 (0.15)	0.26 (0.11)	0.18 (0.09)	0.17 (0.23)
	$spec.$	0.74 (0.11)	0.77 (0.09)	0.84 (0.04)	0.85 (0.05)	0.83 (0.21)
ZH_d	ccr	**0.91** (0.02)	**0.90** (0.02)	**0.89** (0.03)	**0.88** (0.03)	**0.85** (0.03)
	$E_t(ccr)$	0.98 (0.02)	0.96 (0.02)	0.97 (0.02)	0.97 (0.02)	0.94 (0.03)
	$sens.$	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	1.00 (0.01)
	$spec.$	0.67 (0.09)	0.70 (0.08)	0.82 (0.07)	0.84 (0.06)	0.93 (0.03)

Figure 5 displays the distribution of the number of non-zero coefficients for each group that was included in the final model, which is revealing in several ways. Firstly, it indicates that PCA finds groups of similar lengths, and secondly, the selected model is sparse at both the group and the variable levels.

In an attempt to discover the biological and genetic meaning in the model selected by our methodology, we ran DAVID (Huang et al., 2008b,a) to detect enriched functional-related gene groups. The clustering and functional annotation was performed using the default analysis options, and the role of the potential multiple testing effect was considered using the false discovery rate (FDR).

We observed just two remarkable families of pathways after the gene enrichment analysis: the homeobox-related and the oxidative phosphorylation pathways. They are both involved in the mechanism of action of docetaxel and carboplatin in response to the provided treatment.

Figure 3: Sorted *Importance indexes*, according to the criterion given in (9), and after running Algorithm 2 with $R = 200$. The cutoff value was set to $K = \lceil \sqrt{N/2} \rceil = 7$, as described in Section 2.5.

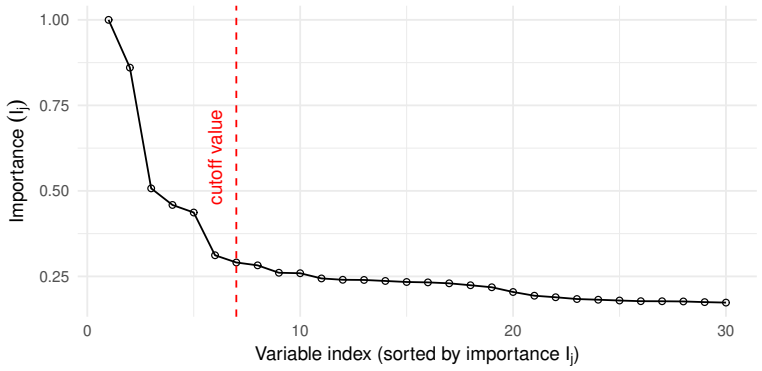
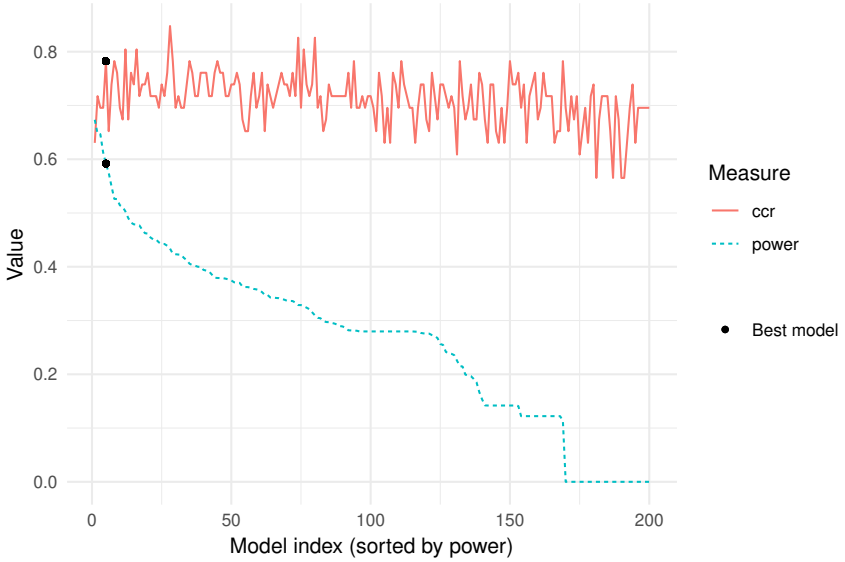


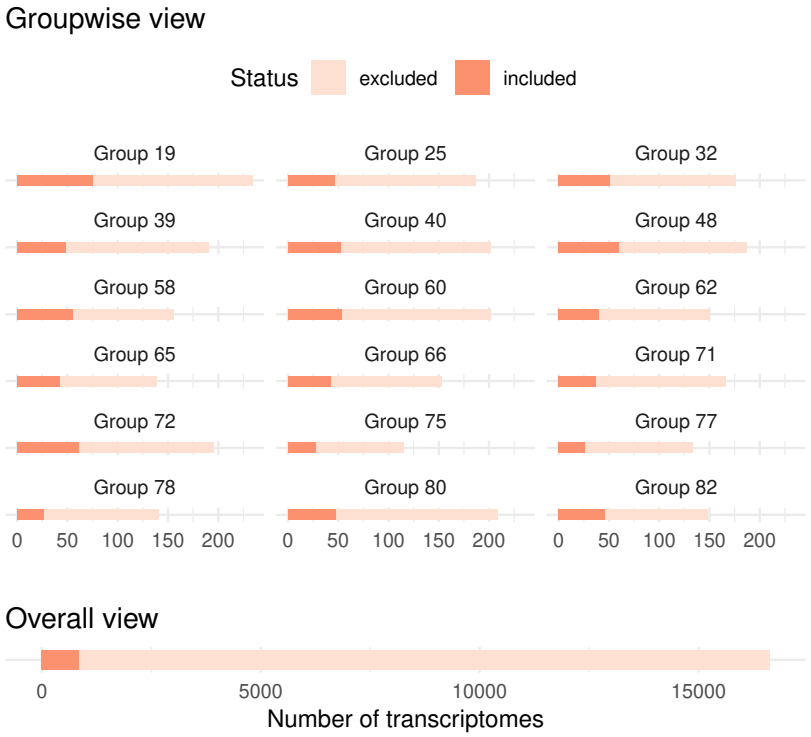
Figure 4: Power index (10), measured in $R = 200$ models, in decreasing order, with the corresponding correct classification rate (ccr) of each model in the validation sample.



The homeobox genes have been proposed to be involved in mechanisms of resistance to taxane-based oncologic treatments in ovarian and prostate cancer (Li et al., 2014; Hanrahan et al., 2017; Marín-Aguilera et al., 2014; Puhr et al., 2012). Docetaxel hyper-stabilizes the microtube structure, irreversibly blocking the cytoskeleton function in the mitotic process and intracellular transport. In addition, this drug induces programmed cell death (Wishart et al., 2017).

On the other hand, carboplatin attaches alkyl groups to DNA bases resulting in fragmentation by repair enzymes when trying to repair it. It also inducts to mutations due to nucleotide despairing and generates DNA cross-links that affects the transcription process (Wishart et al., 2017). The development of resistance to platinum-based schemes of chemotherapy is a common feature. Several studies demonstrate that dysfunctions in mitochondrial processes, in conjunction with the mentioned mechanism of action, can contribute to develop the phenotypes associated with resistance (Matassa et al., 2016; Dai et al., 2010; Chappell et al., 2012; Marrache et al., 2014; Belotte et al., 2014; McAdam et al., 2016).

Figure 5: Number of included variables in the final model, by groups (top) and total (bottom). There were included 18 out of 82 groups.



5 Conclusions

The present study introduces a methodology to deal with the variable selection problem in the high dimensional set-up. It can be seen as an extension of the sparse-group lasso regularization method, without the dependencies on both the hyper-parameters and the groups. There are several critical components in this approach,

- A clustering on the variables, based on PCA, makes it possible to work with an arbitrarily large number of variables, without specifying groups apriori.
- The iterative sparse group lasso removes the dependence on the hyper-parameters of the sparse group lasso, but it is sensible to the train/validate sample partitions. This problem has been solved running the algorithm for a large number of different train/validate sample partitions (Algorithm 2).
- The correct classification rate of each model in its respective validation sample is stored. Notice that this is an overestimation of the true correct classification rate on future observations, and the highest validation rate does not imply the best model.
- The *importance index* weights the variables, based on the correct classification rate of the models that include them.
- The *power index* weights the models, based on the *importance* of the variables they include.

This methodology was tested on a sample of TNBC patients, trying to reveal the genetic profile associated with resistance to the treatment of interest. The literature studies mentioned in Section 4

provide a rationale supporting the potential predictive value of the two gene pathways identified in our study (the homeobox-related and the oxidative phosphorylation pathways). In order to validate these results, we are testing the model in a new cohort of TNBC patients from the same clinical trial.

Future studies should examine other strategies to group the variables, as discussed in Section 2.3, based on supervised algorithms as well as unsupervised ones.

Acknowledgements

Simulations in Sections 3 and 4 have been carried out in Uranus, a supercomputer cluster located at Universidad Carlos III de Madrid and funded jointly by EU-FEDER funds and by the Spanish Government via the National Projects No. UNC313-4E-2361, No. ENE2009-12213- C03-03, No. ENE2012-33219 and No. ENE2015-68265-P.

References

Azevedo Costa, M., T. de Souza Rodrigues, A. G. F. da Costa, R. Natowicz, and A. Pádua Braga (2017). Sequential selection of variables using short permutation procedures and multiple adjustments: An application to genomic data. *Statistical methods in medical research 26*(2), 997–1020.

Belotte, J., N. M. Fletcher, A. O. Awonuga, M. Alexis, H. M. Abu-Soud, M. G. Saed, M. P. Diamond, and G. M. Saed (2014). The role of oxidative stress in the development of cisplatin resistance in epithelial ovarian cancer. *Reproductive sciences 21*(4), 503–508.

Chang, J. C., E. C. Wooten, A. Tsimelzon, S. G. Hilsenbeck, M. C. Gutierrez, R. Elledge, S. Mohsin, C. K. Osborne, G. C. Chamness, D. C. Allred, et al. (2003). Gene expression profiling for the prediction of therapeutic response to docetaxel in patients with breast cancer. *The Lancet 362*(9381), 362–369.

Chappell, N. P., P.-n. Teng, B. L. Hood, G. Wang, K. M. Darcy, C. A. Hamilton, G. L. Maxwell, and T. P. Conrads (2012). Mitochondrial proteomic analysis of cisplatin resistance in ovarian cancer. *Journal of proteome research 11*(9), 4605–4614.

Cortazar, P., L. Zhang, M. Untch, K. Mehta, J. P. Costantino, N. Wolmark, H. Bonnefoi, D. Cameron, L. Gianni, P. Valagussa, et al. (2014). Pathological complete response and long-term clinical benefit in breast cancer: the ctneobc pooled analysis. *The Lancet 384*(9938), 164–172.

Dai, Z., J. Yin, H. He, W. Li, C. Hou, X. Qian, N. Mao, and L. Pan (2010). Mitochondrial comparative proteomics of human ovarian cancer cells and their platinum-resistant sublines. *Proteomics 10*(21), 3789–3799.

De Los Campos, G., D. Gianola, and D. B. Allison (2010). Predicting genetic predisposition in humans: the promise of whole-genome markers. *Nature Reviews Genetics 11*(12), 880.

Dent, R., M. Trudeau, K. I. Pritchard, W. M. Hanna, H. K. Kahn, C. A. Sawka, L. A. Lickley, E. Rawlinson, P. Sun, and S. A. Narod (2007). Triple-negative breast cancer: clinical features and patterns of recurrence. *Clinical cancer research 13*(15), 4429–4434.

Ferlay, J., I. Soerjomataram, M. Ervik, R. Dikshit, S. Eser, C. Mathers, M. Rebelo, D. Parkin, D. Forman, and F. Bray (2014). Globocan 2012 v1. 0, cancer incidence and mortality worldwide: Iarc cancerbase no. 11. lyon, france: International agency for research on cancer; 2013.

Hanrahan, K., A. O'neill, M. Prencipe, J. Bugler, L. Murphy, A. Fabre, M. Puhr, Z. Culig, K. Murphy, and R. W. Watson (2017). The role of epithelial–mesenchymal transition drivers zeb1 and zeb2 in mediating docetaxel-resistant prostate cancer. *Molecular oncology 11*(3), 251–265.

Hatzis, C., L. Pusztai, V. Valero, D. J. Booser, L. Esserman, A. Lluch, T. Vidaurre, F. Holmes, E. Souchon, H. Wang, et al. (2011). A genomic predictor of response and survival following taxane-anthracycline chemotherapy for invasive breast cancer. *Jama 305*(18), 1873–1881.

Huang, D. W., B. T. Sherman, and R. A. Lempicki (2008a). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic acids research 37*(1), 1–13.

Huang, D. W., B. T. Sherman, and R. A. Lempicki (2008b). Systematic and integrative analysis of large gene lists using david bioinformatics resources. *Nature protocols 4*(1), 44.

Laria, J. C., M. Carmen Aguilera-Morillo, and R. E. Lillo (2019). An iterative sparse-group lasso. *Journal of Computational and Graphical Statistics*, 1–10.

Li, J., Y. Zhang, Y. Gao, Y. Cui, H. Liu, M. Li, and Y. Tian (2014). Downregulation of hnf1 homeobox b is associated with drug resistance in ovarian cancer. *Oncology reports 32*(3), 979–988.

Lupski, J. R., J. W. Belmont, E. Boerwinkle, and R. A. Gibbs (2011). Clan genomics and the complex architecture of human disease. *Cell 147*(1), 32–43.

Marín-Aguilera, M., J. Codony-Servat, Ò. Reig, J. J. Lozano, P. L. Fernández, M. V. Pereira, N. Jiménez, M. Donovan, P. Puig, L. Mengual, et al. (2014). Epithelial-to-mesenchymal transition mediates docetaxel resistance and high risk of relapse in prostate cancer. *Molecular cancer therapeutics*.

Marrache, S., R. K. Pathak, and S. Dhar (2014). Detouring of cisplatin to access mitochondrial genome for overcoming resistance. *Proceedings of the National Academy of Sciences*, 201405244.

Matassa, D., M. Amoroso, H. Lu, R. Avolio, D. Arzeni, C. Procaccini, D. Faicchia, F. Maddalena, V. Simeon, I. Agliarulo, et al. (2016). Oxidative metabolism drives inflammation-induced platinum resistance in human ovarian cancer. *Cell death and differentiation 23*(9), 1542.

McAdam, E., R. Brem, and P. Karran (2016). Oxidative stress–induced protein damage inhibits dna repair and determines mutation risk and therapeutic efficacy. *Molecular Cancer Research*.

Offit, K. (2011). Personalized medicine: new genomics, old lessons. *Human genetics 130*(1), 3–14.

Puhr, M., J. Hoefer, G. Schäfer, H. H. Erb, S. J. Oh, H. Klocker, I. Heidegger, H. Neuwirt, and Z. Culig (2012). Epithelial-to-mesenchymal transition leads to docetaxel resistance in prostate cancer and is mediated by reduced expression of mir-200c and mir-205. *The American journal of pathology 181*(6), 2188–2201.

Sharma, P., S. López-Tarruella, J. A. Garcia-Saenz, Q. J. Khan, H. Gomez, A. Prat, F. Moreno, Y. Jerez-Gilarranz, A. Barnadas, A. C. Picornell, et al. (2018). Pathological response and survival in triple-negative breast cancer following neoadjuvant carboplatin plus docetaxel. *Clinical Cancer Research*, clincanres–0585.

Sharma, P., S. López-Tarruella, J. A. García-Saenz, C. Ward, C. Connor, H. L. Gómez, A. Prat, F. Moreno, Y. Jerez-Gilarranz, A. Barnadas, et al. (2016). Efficacy of neoadjuvant carboplatin plus docetaxel in triple negative breast cancer: Combined analysis of two cohorts. *Clinical Cancer Research*, clincanres–0162.

Simon, N., J. Friedman, T. Hastie, and R. Tibshirani (2013). A sparse-group lasso. *Journal of computational and graphical statistics 22*(2), 231–245.

Symmans, W. F., C. Wei, R. Gould, X. Yu, Y. Zhang, M. Liu, A. Walls, A. Bousamra, M. Ramineni, B. Sinn, et al. (2017). Long-term prognostic risk after neoadjuvant chemotherapy associated with residual cancer burden and breast cancer subtype. *Journal of clinical oncology: official journal of the American Society of Clinical Oncology 35*(10), 1049–1060.

Szymczak, S., J. M. Biernacka, H. J. Cordell, O. González-Recio, I. R. König, H. Zhang, and Y. V. Sun (2009). Machine learning in genome-wide association studies. *Genetic epidemiology 33*(S1).

Tabchy, A., V. Valero, T. Vidaurre, A. Lluch, H. L. Gomez, M. Martin, Y. Qi, L. J. Barajas-Figueroa, E. A. Souchon, C. Coutant, et al. (2010). Evaluation of a 30-gene paclitaxel, fluorouracil, doxorubicin and cyclophosphamide chemotherapy response predictor in a multicenter randomized trial in breast cancer. *Clinical Cancer Research*, clincanres–1265.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological) 58*(1), 267–288.

Wishart, D. S., Y. D. Feunang, A. C. Guo, E. J. Lo, A. Marcu, J. R. Grant, T. Sajed, D. Johnson, C. Li, Z. Sayeeda, et al. (2017). Drugbank 5.0: a major update to the drugbank database for 2018. *Nucleic acids research 46*(D1), D1074–D1082.

Yuan, M. and Y. Lin (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 68*(1), 49–67.

Zou, H. and T. Hastie (2003). Regression shrinkage and selection via the elastic net, with applications to microarrays. *Journal of the Royal Statistical Society: Series B. v67*, 301–320.

Zou, H. and T. Hastie (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67*(2), 301–320.